Stochastic Processes and their Applications 26 (1987) 123–140
North-Holland

AN EXPECTED AVERAGE REWARD CRITERION

K.-J. BIERTH

Institut für Angewandte Mathematik, Universität Bonn, 5300 Bonn, FR Germany

Received 31 December 1986
Revised 27 July 1987

In the present paper the expected average reward criterion is considered instead of the average expected reward criterion commonly used in stochastic dynamic programming. This new criterion seems to be more natural and yields stronger results. In addition to the theory of Markov chains, the theory of martingales will be used. This paper is concerned with Markov decision models with finite state space, arbitrary action space and bounded reward functions. In such a model there is always available a Markov policy which almost maximizes the average reward over a unit of time for different criteria. If the action space is a compact metric space there is even a stationary policy with the same property; further if a stationary policy is optimal for one criterion then this policy is optimal for all average reward criteria. Thus the paper solves some problems posed by Demko and Hill (1984).

AMS Subject Classifications: 60J10, 90C47.

Markov decision models * dynamic programming * average reward criteria (ε-)optimal policies * Markov policies * stationary policies

1. Introduction and main results

Consider a Markovian decision model [MDM] defined by $(I, A(i), r(i, a), p_{ij}(a))$. Such a model describes a dynamic system which is observed by a decision maker at discrete time points $t = 0, 1, 2, \ldots$ to be in one of the states of the *state space* I. We assume that I is finite. If at time point t the system is observed in state i, the decision maker controls the system by choosing an action from the *action space* $A(i)$, the set of available actions, which is independent of t. If action a is chosen in state i, then the following happens, independently of the history of the process:

- a *reward* $r(i, a)$ is earned immediately where $r(i, a)$ is a bounded function.
- the system will be in state j at the next time point with *transition* probability $p_{ij}(a)$.

A *decision rule* π_t at time t is a function that assigns to each action the probability of that action being taken at time t; in general, it may depend on all realized states up to and including time t and all realized actions up to time t. A *policy* π is a sequence of decision rules $\pi = (\pi_0, \pi_1, \pi_2, \ldots)$. If all decision rules depend only on the present state and the time point then this policy is called a *Markov policy*. A policy is said to be *stationary* and *deterministic* if all decision rules are identical

and nonrandomized. Hence a stationary and deterministic policy is completely described by a mapping $f: I \to A := \bigcup_{i \in I} A(i)$ such that $f(i) \in A(i)$ for each $i \in I$ and denoted by f^∞. Under each stationary policy f^∞ the sequence of states forms a Markov chain with stationary transition probabilities. Let Δ be the set of all policies. Set $F := \bigtimes_{i \in I} A(i)$ and identify F with the class of all deterministic stationary policies. α is a σ-algebra on A such that $r(i, a)$ and $p_{ij}(a)$ are measurable in a on $A(i)$. For a fixed initial state $i_0 = i$ each policy π defines a probability measure $P_{i,\pi}$ on the trajectory space

$$(\Omega, \gamma) = \left(\bigtimes_0^\infty (I \times A), \bigotimes_0^\infty (\mathcal{P}(I) \otimes \alpha) \right)$$

and a stochastic process $\{(X_n, A_n), n \geq 0\}$ where X_n and A_n describe the state and action at time n, respectively. We shall denote by $E_{i,\pi}$ the expectation with respect to the measure $P_{i,\pi}$. We write, for $f \in F$,

$$P(f) := (p_{ij}(f(i)); i, j \in I) \quad \text{(transition matrix)}$$

$$r(i, f) := r(i, f(i))$$

and

$$P^*(f) := \lim_{n \to \infty} \frac{1}{n} \sum_{t=0}^{n-1} (P(f))^t \quad \text{(Cesaro-limit, cp. [1])}$$

We also agree to identify real-valued functions on I with the corresponding column vectors and an (in)equality connecting two vectors means that the corresponding relation is fulfilled coordinatewise. By $[B]$ we denote the closure of a set B. If we consider another MDM′ then every quantity is denoted by "′".

Let $\|\mu\|_v$ denote the total-variation of a measure μ. If $L = (l_{ij})$ is a matrix of order $n \times m$ and $r = (r_i)$ a vector of dimension n then we denote

$$\|L\| = \sum_{i=1}^{n} \sum_{j=1}^{m} |l_{ij}|, \quad \|r\| = \sum_{i=1}^{n} |r_i|.$$

The aim of control is to maximize the average reward over a unit of time. This is done for several different criteria. Usually the following two criteria $\underline{V}$ and $\bar{V}$ are considered (cp. [1, 4, 8–13, 15, 17, 19]). For all $i \in I$, $\pi \in \Delta$ the average expected reward over a unit of time is defined by

$$\bar{V}(i, \pi) := \overline{\lim_{n \to \infty}} E_{i,\pi} \left[\frac{1}{n} \sum_{t=0}^{n-1} r(X_t, A_t) \right]$$

or

$$\underline{V}(i, \pi) := \underline{\lim_{n \to \infty}} E_{i,\pi} \left[\frac{1}{n} \sum_{t=0}^{n-1} r(X_t, A_t) \right].$$

Dubins and Savage in [6] considered as the gain function the expectation $E \overline{\lim} r(X_n, A_n)$ where only the top rewards are counted. This is not so if we consider

the expected average reward over a unit of time which is defined for all $i \in I$, $\pi \in \Delta$ by

$$\bar{U}(i, \pi) := E_{i,\pi}\left[\overline{\lim_{n\to\infty}} \frac{1}{n} \sum_{t=0}^{n-1} r(X_t, A_t)\right]$$

or

$$\underline{U}(i, \pi) := E_{i,\pi}\left[\underline{\lim_{n\to\infty}} \frac{1}{n} \sum_{t=0}^{n-1} r(X_t, A_t)\right].$$

These criteria are close to the usual criteria. V. Mandl [18] first used the criterion $\bar{U}$ for the irreducible case. The criteria $\underline{U}$ and $\bar{U}$ are in some sense the more natural criteria and the known results of dynamic programming like the optimality equation and optimality principle for general gain functions like $Eg(X_0, A_0, X_1, \ldots)$ can be used. For the criteria $\bar{V}$ and $\underline{V}$ we need the theory of Markov chains but for the criteria $\bar{U}$ and $\underline{U}$ we have to use some results of the martingale theory (cp. [18]). Since for any $f \in F$ the limit for the criterion V exists we write $V(f^\infty)$ instead of $\underline{V}(f^\infty)$ or $\bar{V}(f^\infty)$.

Set

$$\underline{V}(i) := \sup_{\pi\in\Delta} \underline{V}(i, \pi), \qquad \bar{V}(i) := \sup_{\pi\in\Delta} \bar{V}(i, \pi)$$

and

$$\underline{U}(i) := \sup_{\pi\in\Delta} \underline{U}(i, \pi), \qquad \bar{U}(i) := \sup_{\pi\in\Delta} \bar{U}(i, \pi)$$

for all $i \in I$.

A policy π is called *ε-optimal*, $\varepsilon \geq 0$, for a criterion $W = \underline{V}, \bar{V}, \underline{U}, \bar{U}$ if $W(\pi) \geq W - \varepsilon$ and is called *strongly ε-optimal*, $\varepsilon \geq 0$, for the criterion $\bar{V}$ or $\bar{U}$ if $\underline{V}(\pi) \geq \bar{V} - \varepsilon$ or $\underline{U}(\pi) \geq \bar{U} - \varepsilon$. A (strongly 0-optimal policy is called an (strongly) *optimal* policy.

In addition we need the value functions for the class of stationary policies

$$V^s(i) := \sup_{f\in F} V(i, f^\infty) \quad \text{and} \quad U^s(i) := \sup_{f\in F} U(i, f^\infty)$$

for all $i \in I$. Set

$$K := \sup_{i\in I} \sup_{a\in A(i)} |r(i, a)| < \infty. \tag{1.1}$$

We also consider the *β-discounted reward* under a policy π, which is defined by

$$V^\beta(i, \pi) := \sum_{t=0}^{\infty} \beta^n E_{i,\pi}[r(X_t, A_t)] \quad \text{for all } i \in I.$$

Set

$$V^\beta(i) := \sup_{\pi\in\Delta} V^\beta(i, \pi) \quad \text{for all } i \in I.$$

Very important for the proofs in this paper is the existence of bounded functions g and h for a fixed $\varepsilon \geq 0$ such that

$$\sup_{f \in F} P(f)g = g$$

and

$$\sup_{f \in F}\{r(f) + P(f)h\} \leq h + g + \varepsilon.$$

The first equation is denoted by "*first optimality equation*" and the last inequality by "*ε-optimality inequality*".

In Fainberg's paper [11] a summary of theorems on the existence of optimal and ε-optimal policies, depending on the properties of the state space and the action space, is given for the criteria V. It is well known that if the state and action space is finite there exists an (strongly) optimal stationary policy for the criteria $\underline{V}$ and $\bar{V}$; there even exists a bounded solution of the optimality equation (see [8, 9]). But if the action space is compact or the state space is countable there may not exist an optimal policy (see [8, 11]). In a series of papers ([8, 10, 11], and others) sufficient conditions are investigated for the existence of optimal stationary and ε-optimal policies. It was shown in [4, 8] that there exists an "almost" optimal stationary policy for the criterion $\underline{V}$ if the sets $A(i)$ are compact. Fainberg [10] extended this result to the criterion $\bar{V}$. In [4, 8, 10], and [11] also reward functions bounded only from above with values in $[-\infty, \infty)$ are considered. For a finite state space it is natural to consider bounded reward functions. In [4] and [11] examples were cited showing that in the case where the state space is finite and the action space is an arbitrary set, there may not exist an ε-optimal randomized stationary policy but there exists an ε-optimal deterministic Markov policy.

Demko and Hill [5] considered the criteria $\underline{U}$ and $\bar{U}$ but only for a special gain function.

The aim is to extend some results, which we know for the criteria $\underline{V}$ and $\bar{V}$ to the criteria $\underline{U}$ and $\bar{U}$. We are able to simplify the proofs in the literature considerably and to show that all results for the criteria $\underline{V}$ and $\bar{V}$ can be carried over to the other criteria $\underline{U}$ and $\bar{U}$.

Under the following assumption:

Assumption (A) (cp. [4, 8, 10]). For any $i \in I$ we have

(i) $A(i)$ is a compact metric space,

(ii) $r(i, a)$ is upper semicontinuous in a,

(iii) $p_{ij}(a)$ is continuous in a for all $j \in I$,

we are able to prove that for any $\varepsilon > 0$ there exists some deterministic stationary policy which is (strongly) ε-optimal for any of the four criteria. Without assumption (A) we can show that there exists some deterministic Markov policy which is (strongly) ε-optimal for any criteria. Moreover we are able to prove that $\underline{U} = \bar{U} = \underline{V} = \bar{V}$.

2. The existence of almost optimal stationary policies

In [8, § 6.3] it is shown that under Assumption (A) there exists for any β, $0<\beta<1$, an optimal stationary policy f_β in the β-discounted model and that the optimality equation is fulfilled, i.e.

$$\begin{aligned} V^\beta(i) &= \sup_{a\in A(i)} \{r(i,a)+\beta \sum_{j\in I} p_{ij}(a) V^\beta(j)\} \\ &= r(i, f_\beta(i))+\beta \sum_{j\in I} p_{ij}(f_\beta(i)) V^\beta(j, f_\beta^\infty)) \qquad (\mathrm{OG}_\beta) \\ &= V^\beta(i, f_\beta^\infty) \quad \text{for all } i\in I. \end{aligned}$$

For any $f\in F$ there exists some bounded function $h: I\to R$ such that

$$r(f)+P(f)h = h+V(f^\infty)$$

(cp. [8, § 7.8]). Mandl showed for the irreducible case that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} r(X_t, A_t) = V(i, f^\infty) \quad P_{i,f^\infty}\text{-a.s.} \quad \text{for all } i\in I. \qquad (2.1)$$

But, if the Markov chain defined by f has more than one closed class, this equation may not be true. Consider the following example:

Example. Consider a MDM $(I, A(i), r(i,a), p_{ij}(a))$ where

$$I=\{c, b, g\}$$

and, for a $f\in F$,

$$p_{ij}(f(i)) = \begin{cases} 1 & \text{for } i=j=b, \\ 1 & \text{for } i=j=g, \\ \frac{1}{2} & \text{for } i=c, j=b, \\ \frac{1}{2} & \text{for } i=c, j=g, \end{cases}$$

$$r=(i, f(i)) = \begin{cases} 0 & \text{for } i=c, b, \\ 1 & \text{for } i=g. \end{cases}$$

Then

$$V(c, f^\infty)=\tfrac{1}{2}$$

and

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} r(X_t, f(X_t)) = 1_{\{g\}}(X_1).$$

So the equation (2.1) is not true. Hence Mandl's method will not work for the present more general situation. We have the following results (which are known from [1] and [8] for the criteria V).

Lemma 2.1. *For any $f \in F$ we have that*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} r(X_t, A_t) = \lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} V(X_t, f^\infty) \quad P_{i,f^\infty}\text{-a.s.} \quad \text{for all } i \in I$$

and

$$\underline{U}(f^\infty) = \bar{U}(f^\infty) = V(f^\infty) = P^*(f)r(f)$$
$$= \lim_{\beta\uparrow 1}(1-\beta)V^\beta(f^\infty).$$

Proof. It is well known (cp. [8, § 7.1ff] and [1]) that for any policy f^∞, $f \in F$

$$V(f^\infty) = \lim_{\beta\uparrow 1}(1-\beta)V^\beta(f^\infty) = P^*(f)r(f). \tag{2.2}$$

Let f^∞, $f \in F$, be a fixed stationary policy. Then there exists a bounded function h satisfying

$$P(f)V(f^\infty) = V(f^\infty) \quad \text{and} \quad r(f) + P(f)h = h + V(f^\infty). \tag{2.3}$$

Set $W := V(f^\infty)$, $P := P(f)$ and $r := r(f)$. First we want to show that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} W(X_t) \text{ exists } P_{i,f^\infty}\text{-a.s. for all } i \in I.$$

From (2.3) it follows that

$$E_{i,f^\infty}[W(X_{t+1}) \mid X_0, X_1, \ldots, X_t] = W(X_t) \quad P_{i,f^\infty}\text{-a.s.}$$

for all $t \in \mathbb{N}$ and so $W(X_n)$ is a martingale and applying Theorem 7.4.3 in [3] it follows that $W(X_n)$ converges towards a random variable W^∞ P_{i,f^∞}-a.s. So we have that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} W(X_t) = W^\infty \quad P_{i,f^\infty}\text{-a.s.} \quad \text{for all } i \in I. \tag{2.4}$$

Now we will show that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} r(X_t) \text{ exists } P_{i,f^\infty}\text{-a.s.}$$

Set

$$Y_n := r(X_n) + h(X_{n+1}) - h(X_n) - W(X_n) \quad \text{for any } n \in \mathbb{N}_0$$

and

$$M_n := \sum_{j=0}^{n-1} Y_j \quad \text{for all } n \in \mathbb{N}.$$

Then Y_n is $\sigma(X_0, X_1, X_2, \ldots, X_{n+1})$-$\mathcal{B}$ measurable and for any initial state $i \in I$ we have that

$$E_{i,f^\infty}[M_{n+1} \mid (X_0, X_1, \ldots, X_n)] = M_n \quad P_{i,f^\infty}\text{-a.s.}$$

So M_n is a martingale and since $|Y_n| \leq C < \infty$ it follows, applying Theorem 32.1.E in [17], that

$$\lim_{n\to\infty} \frac{1}{n} M_n = 0 \quad P_{i,f^\infty}\text{-a.s.} \quad \text{for all } i \in I.$$

Hence (since h is bounded)

$$0 = \lim_{n\to\infty} \frac{1}{n} \left[\sum_{t=0}^{n-1} r(X_t) + h(X_n) - h(X_0) - \sum_{t=0}^{n-1} W(X_t) \right]$$

$$= \lim_{n\to\infty} \frac{1}{n} \left[\sum_{t=0}^{n-1} r(X_t) - \sum_{t=0}^{n-1} W(X_t) \right] \quad P_{i,f^\infty}\text{-a.s.} \quad \text{for all } i \in I,$$

and with (2.4) we get

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} r(X_t) = \lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} W(X_t) \quad P_{i,f^\infty}\text{-a.s.} \quad \text{for all } i \in I. \tag{2.5}$$

With the dominated convergence theorem (and by the definition of W) we get

$$\underline{U}(f^\infty) = V(f^\infty) = \bar{U}(f^\infty). \qquad \square \tag{2.6}$$

So we write $U(f^\infty)$ instead of $\underline{U}(f^\infty)$ or $\bar{U}(f^\infty)$ and we get that $U^s = V^s$. In a similar way as in [8, § 7.13] Schäl [20] shows that, under Assumption (A), $\lim_{\beta\uparrow 1}(1-\beta)V^\beta$ exists and so we get the following results.

Lemma 2.2. *Under Assumption* (A),

(i) $V^s = \lim_{\beta\uparrow 1}(1-\beta)V^\beta = \underline{V} = \underline{U}$;

(ii) *for any* $\varepsilon > 0$ *there exists some* $f \in F$ *such that*

$$U(f^\infty) \geq \underline{U} - \varepsilon \quad \textit{and} \quad V(f^\infty) \geq \underline{V} - \varepsilon.$$

Proof. (i) In [8, § 7.13] it is shown that

$$V^s = \underline{V} \tag{2.7}$$

and with Lemma 2.1 and Fatou's Lemma we get

$$\underline{U} = \underline{V}. \tag{2.8}$$

In [20] it is shown that

$$\lim_{\beta\uparrow 1}(1-\beta)V^\beta = V^s$$

and so we get (i).

(ii) In [4, 8] and [10] it was already proved that for every $\varepsilon > 0$ there exists some $f \in F$ such that $V(f^\infty) \geq \underline{V} - \varepsilon$ and, with (2.6) and (2.8), (ii) follows. $\square$

This lemma is also true for reward functions bounded only from above. Define

$$g := \underline{V} = \underline{U} = V^s = U^s = \lim_{\beta \uparrow 1}(1-\beta)V^\beta$$

and

$$s := |I|.$$

For the proof of Lemma 2.2 it is very important that I is finite and for the proof of the next lemma that $\lim_{\beta \to 1}(1-\beta)V^\beta$ exists and is equal to g.

Lemma 2.3. *Under assumption* (A) *we have for any* $i \in I$ *that*

(i) $\sup_{a \in A(i)} \sum_{j \in I} p_{ij}(a) g(j) = g(i)$;

(ii) *for any* $\varepsilon > 0$ *there exists some bounded function* h *such that*

$$\sup_{a \in A(i)} \left\{ r(i,a) + \sum_{j \in I} p_{ij}(a) h(j) \right\} \leq g(i) + h(i) + \varepsilon$$

and

$$\sup_{a \in A(i)} \left\{ r(i,a) + \sum_{j \in I} p_{ij}(a)[h(j) - g(j)] \right\} \leq h(i) + \varepsilon.$$

Proof. (i) see [8, § 7.13].

(ii) Choose an $\varepsilon > 0$. From Lemma 2.2 it follows that $\lim_{\beta \uparrow 1}(1-\beta)V^\beta = g.$ and since I is finite we can find a β, $1 > \beta > 0$, such that

$$\|(1-\beta)V^\beta - g\| \leq \varepsilon. \tag{2.9}$$

Set $h := V^\beta$ then

$$\|h\| = \|V^\beta\| \leq \frac{K}{1-\beta} < \infty \tag{2.10}$$

and, for a fixed $i \in I$ and $a \in A(i)$,

$$\begin{aligned} & r(i,a) + \sum_{j \in I} p_{ij}(a) h(j) \\ & = r(i,a) + \beta \sum_{j \in I} p_{ij}(a) V^\beta(j) + \sum_{j \in I} p_{ij}(a)(1-\beta)V^\beta(j) \\ & \leq V^\beta(i) + \sum_{j \in I} p_{ij}(a)[(1-\beta)V^\beta(j) - g(j)] + \sum_{j \in I} p_{ij}(a) g(j) \quad (\mathrm{OG}_\beta) \\ & \leq h(i) + \sum_{j \in I} p_{ij}(a) g(j) + \varepsilon \leq h(i) + g(i) + \varepsilon \quad \text{(see (2.9) and (i)).} \end{aligned}$$

Hence

$$\sup_{a \in A(i)} \left\{ r(i,a) + \sum_{j \in I} p_{ij}(a)[h(j) - g(j)] \right\} \leq h(i) + \varepsilon$$

and

$$\sup_{a\in A(i)}\left\{r(i,a)+\sum_{j\in I}p_{ij}(a)h(j)\right\}\leq h(i)+g(i)+\varepsilon$$

for all $i\in I$. □

The following lemma was proved for the criteria V in [8, § 7.9].

Lemma 2.4. *If for an $\varepsilon>0$ there exist bounded functions $\tilde{g}$ and h, $\tilde{g}, h: I\to R$, such that, for any $i\in I$,*

(i) $\sup_{a\in A(i)}\{\sum_{j\in I}p_{ij}(a)\tilde{g}(j)\}=\tilde{g}(i)$, *and*

(ii) $\sup_{a\in A(i)}\{r(i,a)+\sum_{j\in I}p_{ij}(a)h(j)\}\leq h(i)+\tilde{g}(i)+\varepsilon$

then we get

$$\bar{U}\leq\tilde{g}+\varepsilon.$$

Proof. Set

$$Z(i,a):=r(i,a)+\sum_{j\in I}p_{ij}(a)h(j)-h(i)-\tilde{g}(i)$$

for all $a\in A(i)$, $i\in I$; then it follows from (ii) that

$$\sup_{a\in A(i)}Z(i,a)\leq\varepsilon\quad\text{for all } i\in I. \tag{2.11}$$

Define

$$Y_n:=r(X_n,A_n)+h(X_{n+1})-h(X_n)-\tilde{g}(X_n)-Z(X_n,A_n)\quad\text{for } n\in\mathbb{N}_0$$

and

$$M_n:=\sum_{m=0}^{n-1}Y_m\quad\text{for all } n\in\mathbb{N}.$$

For all $n\in\mathbb{N}_0$ Y_n is $\sigma(X_0,A_0,\ldots,A_{n+1})$-$\mathcal{B}$ measurable. Now we choose a fixed initial state i and a fixed policy π. Then we get, from (ii),

$$E_{i,\pi}[Y_n\,|\,X_0,A_0,X_1,\ldots,A_n)]=0\ \ P_{i,\pi}\text{-a.s.}$$

and so M_n is a martingale. Since $|Y_n|\leq C<\infty$ for all $n\in\mathbb{N}_0$ we get, applying Theorem 32.1.E in [17], that

$$\lim_{n\to\infty}\frac{1}{n}M_n=0\ \ P_{i,\pi}\text{-a.s.}$$

and so, with (2.11),

$$0=\lim_{n\to\infty}\frac{1}{n}\left[\sum_{t=0}^{n-1}r(X_t,A_t)+h(X_n)-h(X_0)-\sum_{t=0}^{n-1}\tilde{g}(X_t)-\sum_{t=0}^{n-1}Z(X_t,A_t)\right]$$

$$\geq\overline{\lim_{n\to\infty}}\frac{1}{n}\sum_{t=0}^{n-1}r(X_t,A_t)-\overline{\lim_{n\to\infty}}\frac{1}{n}\sum_{t=0}^{n-1}\tilde{g}(X_t)-\varepsilon\ \ P_{i,\pi}\text{-a.s.}$$

Hence

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \sum_{t=0}^{n-1} r(X_t, A_t) \leq \overline{\lim_{n\to\infty}} \frac{1}{n} \sum_{t=0}^{n-1} \tilde{g}(X_t) + \varepsilon \quad P_{i,\pi}\text{-a.s.} \tag{2.12}$$

and

$$\bar{U}(i, \pi) \leq E_{i,\pi} \overline{\lim_{n\to\infty}} \frac{1}{n} \sum_{t=0}^{n-1} \tilde{g}(X_t) + \varepsilon. \tag{2.13}$$

From (i) it follows that $E_{i,\pi}[\tilde{g}(X_n)] \leq \tilde{g}(i)$ for all $i \in I$, $n \in \mathbb{N}_0$ and so

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \sum_{j=0}^{n-1} E_{i,\pi}[\tilde{g}(X_j)] \leq \tilde{g}(i). \tag{2.14}$$

Also from (i) it follows that $\tilde{g}(X_n)$ is a supermartingale ($\tilde{g}(X_n)$ is $\sigma(X_0, A_0, \ldots, A_n)$-$\mathcal{B}$ measurable and $E_{i,\pi}[\tilde{g}(X_{n+1}) | X_0, A_0, \ldots, A_n] \leq \tilde{g}(X_n)$ $P_{i,\pi}$-a.s.). Applying Theorem 7.4.2 in [3] we have that $\tilde{g}(X_n)$ converges towards a random variable $V^{i,\pi}$ $P_{i,\pi}$-a.s. ($\sup_{n\in\mathbb{N}_0} E_{i,\pi}|\tilde{g}(X_n)| \leq K$). So we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \tilde{g}(X_t) = V^{i,\pi} \quad P_{i,\pi}\text{-a.s.}$$

Hence

$$\begin{aligned} E_{i,\pi} V^{i,\pi} &= E_{i,\pi} \lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \tilde{g}(X_t) \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} E_{i,\pi}[\tilde{g}(X_t)] \\ &\leq \tilde{g}(i) \quad \text{(dominated convergence and (2.14))}. \end{aligned} \tag{2.15}$$

From (2.13) and (2.15) it follows that $\bar{U}(i, \pi) \leq \tilde{g}(i) + \varepsilon$ and since i and π were arbitrary chosen we get that

$$\bar{U} \leq \tilde{g} + \varepsilon. \qquad \square \tag{2.16}$$

Now we can prove that $\bar{U} = g$ and so we get the following theorem.

Theorem 2.5. *Under Assumption* (A)

(i) $\underline{V} = \bar{V} = \underline{U} = \bar{U} = V^s = U^s = \lim_{\beta\uparrow 1}(1-\beta) V^\beta$ $(= g)$;

(ii) *for any* $\varepsilon > 0$ *there exists some* $f \in F$ *such that*

$$V(f^\infty) \geq \bar{V} - \varepsilon = \underline{V} - \varepsilon \quad \textit{and} \quad U(f^\infty) \geq \bar{U} - \varepsilon = \underline{U} - \varepsilon.$$

Proof. (i) Choose a fixed $\varepsilon > 0$. Then by Lemma 2.3 there exists a bounded function h such that

$$\sup_{a\in A(i)} \sum_{j\in I} p_{ij}(a) g(j) = g(i) \tag{2.18}$$

and

$$\sup_{a\in A(i)}\left\{r(i,a)+\sum_{j\in I}p_{ij}(a)h(j)\right\}\leqslant h(i)+g(i)+\varepsilon \quad \text{for all } i\in I. \tag{2.19}$$

From (2.18) and (2.19) it follows with Lemma 2.4(i) that

$$\bar{U}\leqslant g+\varepsilon. \tag{2.20}$$

Since ε was arbitrary chosen and $g\leqslant \bar{V}\leqslant \bar{U}$ (Fatou's Lemma) we have

$$\bar{V}=\bar{U}=g. \tag{2.21}$$

(ii) follows from (2.21) and Lemmas 2.2(ii) and 2.1. □

In [10, Lemma 9] it was already proved that

$$\underline{V}=\bar{V}=V^s=\lim_{\beta\uparrow 1}(1-\beta)V^\beta$$

and that there exists for any $\varepsilon>0$ some $f\in F$ such that

$$\bar{V}(f^\infty)=\underline{V}(f^\infty)\geqslant \bar{V}-\varepsilon=\underline{V}-\varepsilon.$$

We get these results in an easier way than in [10]. So far we proved the existence of "almost optimal" stationary policies. It is known that even under Assumption (A) there may not exist an optimal policy (cp. [8, § 7.8]) but it follows from Theorem 2.5(i) that if there exists an optimal stationary policy for one of the four criteria then this policy is (strongly) optimal for all criteria.

Corollary 2.6. *There exists an (strongly) optimal stationary policy for every criterion if one of the following assumptions holds:*

(i) $A(i)$ *is finite for any* $i\in I$ (cp. [9]);
(ii) $|I|=2$ (cp. [9]);
(iii) $P^*(\cdot)$ *is continuous on* F (cp. [15, *Theorem* 10.4]);
(iv) *there exists an optimal policy for one of the criteria* $\underline{V}$, $\bar{V}$, $\bar{U}$, $\underline{U}$;
(v) *each Markov Chain defined by* f *has a single ergodic class* (cp. [8, 9]);
(vi) *the sets*

$$Y_i=\{y\in\mathbb{R}^s \mid \text{there exists a, } a\in A(i), \text{ with } y_i=p_{i\cdot}(a)\}$$

have a finite number of extreme points (cp. [9]).

3. The existence of almost optimal Markov policies

If we do not have Assumption (A) in [4] an example is given which shows that there exists no ε-optimal stationary policy for an $\varepsilon>0$ (so $V^s<V$). But as we will show below there exists an ε-optimal Markov policy for any $\varepsilon>0$ and any criterion. For that purpose we embed the model MDM in a model MDM′ which fulfills Assumption (A).

Definition. A MDM defined by $(I, A'(i), r'(i, a), p'_{ij}(a))$ is a *representation* of a MDM defined by $(I, A(i), r(i, a), p_{ij}(a))$ if for every $i \in I$ there exists a surjective mapping ψ_i of the set $A(i)$ onto the set $A'(i)$ such that

$$p_{ij}(a) = p'_{ij}(\psi_i(a))$$

and

$$r(i, a) = r'(i, \psi_i(a)) \quad \text{for all } a \in A(i),\ i \in I.$$

Remark. It is easy to see that $U' = U$. For any $f' \in F'$ there exists some $f \in F$ such that $U'(f'^{\infty}) = U(f^{\infty})$ and conversely. A similar definition to the above and a lemma similar to the following were given in [11].

Lemma 3.1. *For any* MDM $(I, A(i), r(i, a), p_{ij}(a))$ *there exists a representation* $(I, A'(i), r'(i, a), p'_{ij}(a))$ *such that*

- $[A'(i)]$ *is a compact metric set for any* $i \in I$
- $r'(i, a)$ *and* $p'_{ij}(a)$ *are uniformly continuous on* $A'(i)$ *for all* $i \in I$.

Proof. For any $i \in I$ define a mapping ψ_i of the set $A(i)$ onto R^{s+1} by

$$\psi_i(a) := (r(i, a), p_{i1}(a) \cdots p_{is}(a))$$

and set

$$A'(i) := \Psi_i(A(i)) \subset \mathbb{R}^{s+1},$$

$$r'(i, u) := u_1 \quad \text{(projecting from } \mathbb{R}^{s+1} \text{ to } \mathbb{R}),$$

$$p'_{i.}(u) := (u_2 \cdots u_{s+1}) \quad \text{(projecting from } \mathbb{R}^{s+1} \text{ to } \mathbb{R}^s).$$

The MDM' $(I, A'(i), r'(i, a), p'_{ij}(a))$ is a representation of the model MDM. By construction we can see that $A'(i)$ is bounded hence $[A'(i)]$ is compact for all $i \in I$. Since

$$\| p'_{i.}(u) - p'_{i.}(u') \| \leq \| u - u' \|$$

and

$$\| r'(i, u) - r'(i, u') \| \leq \| u - u' \| \quad \text{for any } u, u' \in A'(i),\ i \in I,$$

we have that $p'_{ij}(\cdot)$ and $r'(i, \cdot)$ are uniformly continuous on $A'(i)$ for any $i \in I$. □

Lemma 3.2 (cp. Lemma 5 in [5]). *Let* $f \in [F]$, $[F]$ *compact, and the functions* $p_{ij}(a)$ *be uniformly continuous on* $A(i)$, *then for any* $\varepsilon > 0$ *there exists some deterministic Markov policy* $\pi \in \Delta$ *such that*

$$\| P_{i,\pi} - P_{i,f^{\infty}} \|_v \leq \varepsilon \quad \forall i \in I.$$

Proof. Let $f \in \bigtimes_{i\in I}[A(i)]$ and $\varepsilon > 0$ be fixed. Choose now a sequence $\{f_n\}_{n\in\mathbb{N}_0}$ of $f_n \in F$ such that (since the functions $p_{ij}(a)$ are uniformly continuous)

$$\begin{aligned} &\text{(a)} \quad f_n(i) \to f(i), \\ &\text{(b)} \quad \| p_{i.}(f_n) - p_{i.}(f) \| \leqslant \varepsilon/(2s)^{n+1} \end{aligned} \tag{3.2}$$

for all $i \in I$. Define a Markov policy π with $\pi_t(i) := f_t(i)\ \forall i \in I$ and

$$\zeta := \bigcup_{T\in\mathcal{P}_0} \left\{ A \times \bigtimes_{l\notin T} I \;\middle|\; A \in \bigotimes_{l\in T} \mathcal{P}(I) \right\}$$

where $\mathcal{P}_0$ is the set of all finite non empty subsets of $\mathbb{N}$. Applying Theorem 1.3.4 in [3] we have that

$$\sigma(\zeta) = \bigotimes_{0}^{\infty} \mathcal{P}(I) \tag{3.3}$$

If $B \in \zeta$ then there exists a $T \in \mathcal{P}_0$, such that

$$B = A \times \bigtimes_{l\notin T} I \quad \text{and} \quad A \in \bigotimes_{l\in T} \mathcal{P}(I).$$

Set $n := \max\{l \mid l \in T\}$, $P^i := P(f_i)$, $P := P(f)$; then we have that

$$\begin{aligned} |P_{i,\pi}(B) - P_{i,f^\infty}(B)| &\leqslant \sum_{i_1\cdots i_n \in I} |p^0_{ii_1} \cdots p^{n-1}_{i_{n-1}i_n} - p_{ii_1} \cdots p_{i_{n-1}i_n}| \\ &\leqslant s^n \max_{i_1\cdots i_n \in I} |p^0_{ii_1} \cdots p^{n-1}_{i_{n-1}i_n} - p_{ii_1} \cdots p_{i_{n-1}i_n}| \\ &\leqslant \frac{\varepsilon}{2^{n+1}} \quad \text{(see (3.2))}. \end{aligned}$$

So we have for all $B \in \zeta$ that

$$|P_{i,\pi}(B) - P_{i,f^\infty}(B)| \leqslant \frac{\varepsilon}{2^2}. \tag{3.4}$$

Let μ be a probability measure on (Ω, γ) then it follows from Theorem 8.1.1 in [2] and (3.3) that for any $\delta > 0$, and $A \in \gamma$ there exists some $B \in \zeta$ such that $\mu(A \Delta B) < \delta$. Choose

$$\mu := \tfrac{1}{2}(P_{i,\pi} + P_{i,f^\infty}) \quad \text{and} \quad \delta := \frac{\varepsilon}{16};$$

then we have that

$$P_{i,\pi}(A\Delta B) + P_{i,f^\infty}(A\Delta B) < \frac{\varepsilon}{8},$$

$$P_{i,\pi}(A\Delta B) < \frac{\varepsilon}{8},$$

and

$$P_{i,f^\infty}(A \Delta B) < \frac{\varepsilon}{8}.$$

Hence

$$|P_{i,\pi}(A) - P_{i,\pi}(B)| < \frac{\varepsilon}{8} \quad \text{and} \quad |P_{i,f^\infty}(B) - P_{i,f^\infty}(A)| < \frac{\varepsilon}{8}. \tag{3.5}$$

From (3.4) and (3.5) it follows that

$$\begin{aligned} &|P_{i,\pi}(A) - P_{i,f^\infty}(A)| \\ &\quad = |P_{i,\pi}(A) - P_{i,\pi}(B) + P_{i,\pi}(B) - P_{i,f^\infty}(B) + P_{i,f^\infty}(B) - P_{i,f^\infty}(A)| \\ &\quad \leq |P_{i,\pi}(A) - P_{i,\pi}(B)| + |P_{i,\pi}(B) - P_{i,f^\infty}(B)| + |P_{i,f^\infty}(B) - P_{i,f^\infty}(A)| \\ &\quad \leq \frac{\varepsilon}{8} + \frac{\varepsilon}{4} + \frac{\varepsilon}{8} = \frac{\varepsilon}{2}. \end{aligned}$$

So for any $A \in \gamma$ we have that

$$|P_{i,\pi}(A) - P_{i,f^\infty}(A)| \leq \frac{\varepsilon}{2} \tag{3.6}$$

and since (see Lemma III.1.5 in [7])

$$\|P_{i,\pi} - P_{i,f^\infty}\|_v \leq 2 \sup_{A \in \gamma} |P_{i,\pi}(A) - P_{i,f^\infty}(A)|$$

we get

$$\|P_{i,\pi} - P_{i,f^\infty}\|_v \leq \varepsilon. \qquad \square$$

Theorem 3.3. (i) $\underline{U} = \bar{U} = \underline{V} = \bar{V}$ $(= g')$.

(ii) *For any $\varepsilon > 0$ there exists some deterministic Markov policy π such that*

$$\underline{U}(\pi) \geq \bar{U} - \varepsilon = \underline{U} - \varepsilon \quad \textit{and} \quad \underline{U}(\pi) = \bar{U}(\pi) = \underline{V}(\pi) = \bar{V}(\pi).$$

Proof. It follows from Lemma 3.1 that without loss of generality, the sets $[A(i)]$ can be assumed compact, while the functions $p_{ij}(a)$ and $r(i, a)$ are uniformly continuous in a. In view of the uniform continuity we shall extend the functions $p_{ij}(a)$ and $r(i, a)$ from $A(i)$ to $[A(i)]$, $i, j \in I$.

Consider now the MDM′ defined by $(I, [A(i)], r(i, a), p_{ij}(a))$. For this model assumption (A) is fulfilled and so by virtue of Theorem 2.5 there exists in the MDM′ for any $\varepsilon > 0$ a stationary policy $\tilde{f}^\infty$, such that $U'(\tilde{f}^\infty) \geq U' - \varepsilon \geq U - \varepsilon$ (since $A(i) \subset [A(i)]$, $i \in I$). Fix an $\varepsilon > 0$ and let f^∞ be an $\varepsilon/2$-optimal policy $(f^\infty \in F')$. Hence

$$U'(f^\infty) \geq U - \frac{\varepsilon}{2}. \tag{3.8}$$

Now we consider a sequence $\{f_n\}_{n\in\mathbb{N}_0}$ such that

$$\text{(a)}\quad f_n(i)\to f(i), f_n(i)\in A(i),$$

$$\text{(b)}\quad \|p_{i\cdot}(f_n(i))-p_{i\cdot}(f(i))\|\leq\frac{1}{(2s)^{n+1}}\cdot\frac{\varepsilon}{4K}, \tag{3.9}$$

$$\text{(c)}\quad \|r(i,f_n(i))-r(i,f(i))\|\leq\frac{\varepsilon}{4s2^{n+1}}\quad\text{for all } i\in I.$$

Let $\{\pi^k\}_{k\in\mathbb{N}_0}$ be a sequence of Markov policies such that

$$\pi_t^k(i)=f_{t+k}(i),\quad i\in I.$$

From (3.9) we get

$$\begin{aligned}
&\|p_{i\cdot}(f)-p_{i\cdot}(f_{t+k})\|\leq\frac{1}{(2s)^{t+1}}\cdot\frac{\varepsilon}{4K(2s)^k},\\
&\|P(f)-P(f_{t+k})\|\leq\frac{1}{(2s)^t}\cdot\frac{\varepsilon}{4K(2s)^k},\\
&\|r(f)-r(f_{t+k})\|\leq\frac{\varepsilon}{2^{t+1}2^{k+2}}.
\end{aligned}\tag{3.10}$$

From Lemma 3.2 and (3.10) it follows that

$$\begin{aligned}
&\|P_{i,\pi k}-P_{i,f^\infty}\|_v\leq\frac{\varepsilon}{4K2^k}\quad\text{for all } i\in I,\ k\in\mathbb{N}_0,\quad\text{and}\\
&\frac{1}{n}\sum_{t=0}^{n-1}|r(X_t,f(X_t))-r(X_t,f_{t+k}(X_t)|\leq\frac{\varepsilon}{2^{k+2}}\quad\text{for all } k,n\in\mathbb{N}_0.
\end{aligned}\tag{3.11}$$

Hence

$$\begin{aligned}
&|\underline{U}(i,\pi^k)-U'(i,f^\infty)|\\
&\quad=\left|E_{i,\pi k}\left[\varliminf_{n\to\infty}\frac{1}{n}\sum_{t=0}^{n-1}r(X_t,f_{t+k}(X_t))\right]-E_{i,\pi k}\left[\varliminf_{n\to\infty}\frac{1}{n}\sum_{t=0}^{n-1}r(X_t,f(X_t))\right]\right.\\
&\qquad\left.+E_{i,\pi k}\left[\varliminf_{n\to\infty}\frac{1}{n}\sum_{t=0}^{n-1}r(X_t,f(X_t))\right]-E_{i,f^\infty}\left[\varliminf_{n\to\infty}\frac{1}{n}\sum_{t=0}^{n-1}r(X_t,f(X_t))\right]\right|\\
&\quad\leq\left|E_{i,\pi k}\left[\varliminf_{n\to\infty}\frac{1}{n}\sum_{t=0}^{n-1}r(X_t,f_{t+k}(X_t))\right]-E_{i,\pi k}\left[\varliminf_{n\to\infty}\frac{1}{n}\sum_{t=0}^{n-1}r(X_t,f(X_t))\right]\right|\\
&\qquad+\left|E_{i,\pi k}\left[\varliminf_{n\to\infty}\frac{1}{n}\sum_{t=0}^{n-1}r(X_t,f(X_t))\right]-E_{i,f^\infty}\left[\varliminf_{n\to\infty}\frac{1}{n}\sum_{t=0}^{n-1}r(X_t,f(X_t))\right]\right|\\
&\quad\leq\frac{\varepsilon}{2^{k+2}}+\|P_{i,\pi k}-P_{i,f^\infty}\|_v\cdot K\quad\text{(see (3.11))}\\
&\quad\leq\frac{\varepsilon}{2^k4}+\frac{\varepsilon}{2^k4}\quad\text{(see (3.11))}\\
&\quad\leq\frac{\varepsilon}{2^{k+1}}\quad\text{for all } i\in I,\ k\in\mathbb{N}_0.
\end{aligned}$$

So we have (see (3.8))

$$\bar{U}(\pi^k) \geq \underline{U}(\pi^k) \geq \bar{U}'(f^\infty) - \frac{\varepsilon}{2^{k+1}} \geq \bar{U} - \frac{\varepsilon}{2^k} \geq \underline{U} - \frac{\varepsilon}{2^k} \quad \text{for all } k \in \mathbb{N}_0 \tag{3.12}$$

and

$$\| \bar{U}(\pi^k) - \underline{U}(\pi^k) \| \leq \frac{s\varepsilon}{2^k}. \tag{3.13}$$

Further we have for a fixed $k \in \mathbb{N}$ that

$$\begin{aligned} \bar{U}(i, \pi^0) &= E_{i,\pi^0}\left[\overline{\lim_{n\to\infty}} \frac{1}{n} \sum_{t=k}^{n-1} r(X_t, f_t(X_t)) \right] \\ &= E_{i,\pi^0}\left[\overline{\lim_{n\to\infty}} \frac{1}{n} \sum_{t=0}^{n-1-k} r(X_{t+k}, f_{t+k}(X_{t+k})) \right] \\ &= \sum_{j\in I} P_{i,\pi^0}(X_k = j) E_{j,\pi^k}\left[\overline{\lim_{n\to\infty}} \frac{1}{n} \sum_{t=0}^{n-1} r(X_t, f_{t+k}(X_t)) \right]. \end{aligned}$$

So

$$\bar{U}(\pi^0) = \prod_{n=0}^{k-1} P(f_n) \bar{U}(\pi^k)$$

and in a similar way we can prove that

$$\underline{U}(\pi^0) = \prod_{n=0}^{k-1} P(f_n) \underline{U}(\pi^k).$$

From this it follows that we have for any $k \in \mathbb{N}$

$$\| \bar{U}(\pi^0) - \underline{U}(\pi^0) \| = \left\| \prod_{n=0}^{k-1} P(f_n)(\bar{U}(\pi^k) - \underline{U}(\pi^k)) \right\| \leq \frac{s \cdot \varepsilon}{2^k}.$$

Hence

$$\bar{U}(\pi^0) = \underline{U}(\pi^0). \tag{3.14}$$

With Fatou's lemma and (3.14) it follows now that ($\pi := \pi^0$)

$$\underline{U}(\pi) = \underline{V}(\pi) = \bar{V}(\pi) = \bar{U}(\pi). \tag{3.15}$$

Since ε was arbitrary we now get from (3.12) and (3.15) that for any $\varepsilon > 0$ there exists some deterministic Markov policy π such that

$$\underline{U}(\pi) \leq \underline{U} \leq \bar{U} \leq \bar{U}(\pi) + \varepsilon \leq \underline{U} + \varepsilon.$$

Hence $\underline{U} = \bar{U}$.

By Fatou's Lemma we have that

$$\underline{U} \leq \underline{V} \leq \bar{V} \leq \bar{U}$$

and so

$$\underline{U} = \underline{V} = \bar{V} = \bar{U}(= g'). \qquad \square$$

Now we get the following Corollary which was already proved in [11] for reward functions bounded only from above.

Corollary 3.4. *For any $\varepsilon > 0$ there exists some deterministic Markov policy π such that*

$$\bar{V}(\pi) = \underline{V}(\pi) \geq \bar{V} - \varepsilon = \underline{V} - \varepsilon.$$

So for any $\varepsilon > 0$ and any criterion there exists some (strongly) ε-optimal deterministic Markov policy. If there exists an (ε-) optimal stationary policy for one criterion, ($\varepsilon \geq 0$) then this policy is also (strongly) (ε-) optimal for any criterion.

Remarks

- If we consider a special reward function $r(i, a) = 1_{\{g\}}$ for some goal $g \in I$ then we get the results of [5] from Theorem 2.5 and 3.3. Moreover we are able to consider as a goal not only a single state g but also a subset of states.
- Demko and Hill had the same idea of proof (construction) as Fainberg [11]. In a similar way we get the result of this chapter. Since we consider only bounded reward functions our proofs are not so complicated as the ones in [10] and [11]. We are also able to extend the results of this paper to models with reward functions bounded only from above in the same way as Fainberg did.
- The martingale ideas were also used in [5].

Acknowledgement

I wish to thank Prof. M. Schäl for his help. I would also like to thank Prof. T. P. Hill for the proof of Lemma 3.2.

References

[1] J.A. Bather, Optimal decision for finite Markov chains, I. Examples, Adv. Appl. Prob. 5 (1973) 328-339.
[2] K.L. Chung, A Course in Probability Theory (Academic Press, New York, 1968).
[3] Y.S. Chow, Probability Theory (Springer-Verlag, New York, 1978).
[4] R. Chitashvili, A controlled finite Markov chain with an arbitrary set of decision, Theory Prob. Appl. 20 (1975) 839-846.
[5] S. Demko and T. Hill, On maximizing the average time at a goal, Stoch. Proc. Appl. 17 (1984) 349-357.
[6] L.E. Dubins and L.J. Savage, How to Gamble if You Must (McGraw-Hill, New York, 1965).
[7] Dunford and J. Schwartz, Linear Operators Part I (Interscience Publishers, New York, 1958).
[8] E. Dynkin and A. Yushkevich, Controlled Markov Processes (Springer-Verlag, Berlin, Heidelberg, New York, 1979).
[9] E.A. Fainberg, On controlled finite state Markov processes with compact control sets, Theory Prob. 20 (1975) 856-862.
[10] E.A. Fainberg, The existence of a stationary ε-optimal policy for a finite Markov chain, Theory Prob. 23 (1978) 297-313.

[11] E.A. Fainberg, An ε-optimal control of a finite Markov chain with an average cost criterion, Theory Prob. 25 (1980) 70-81.
[12] A. Federgruen, P.J. Schweitzer and H.C. Tijms, Denumerable undiscounted semi-Markov decision process with unbounded rewards, Math. of Oper. Res. 8 (1983) 293-313.
[13] A. Federgruen, A. Hordijk and H.C. Tijms, Denumerable state semi-Markov decision processes with unbounded costs, Average cost criterion, Stoch. Proc. Appl. 9 (1979) 223-235.
[14] L. Fisher, S.M. Ross, An example in denumerable decision processes, Ann. Math. Stat. 39 (1968) 674-675.
[15] A. Hordijk, Dynamic programming and Markov potential theory, Mathematical Centre Tracts 51, Amsterdam (1974).
[16] R.A. Howard, Dynamic Programming and Markov Processes (Technology Press, Cambridge, MA, 1960).
[17] M. Loeve, Probability Theory II (D. van Nostrand, Princeton, NJ, 1960).
[18] P. Mandl, Estimation and control in Markov chains, Adv. Appl. Prob. 6 (1974) 40-60.
[19] S.M. Ross, Applied Probability Models with Optimization Applications (Holden-Day, San Francisco, 1970).
[20] M. Schäl, Markoffsche Entscheidungsprozesse, Vorlesungsreihe SFB 72, Institut für Angewandte Mathematik Universität Bonn (1986).